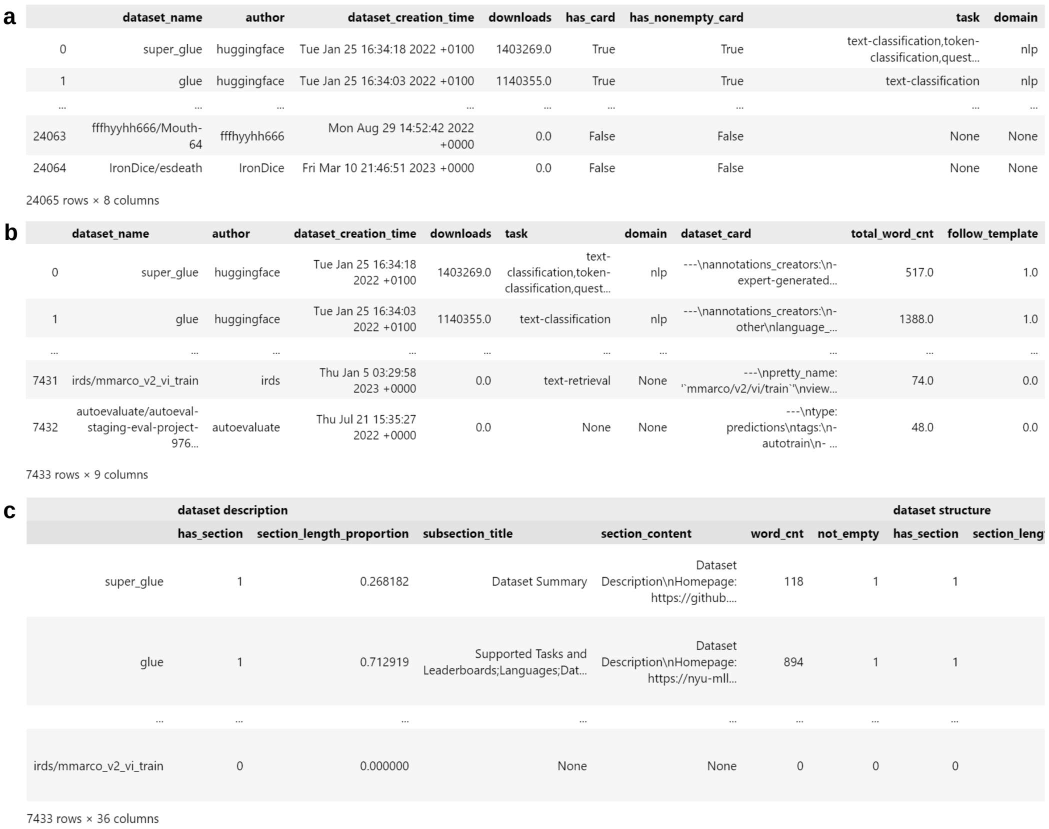
A Code and Data Availability

We have assembled a collection of dataset cards as a community resource, which includes extracted metadata such as the number of downloads and textual analyses. This resource along with our analysis code can be accessed at `https://anonymous.4open.science/r/HuggingFace-Dataset-Card-Analysis`.

The repository comprises two main components: the $Data$ folder and the $Scripts$ folder. The $Data$ folder contains data on 7,433 dataset cards that have been analyzed, along with metadata for each dataset and dataset card. Details about this metadata can be found in **Fig. S1**. The $Scripts$ folder contains the code used to conduct the analysis, which includes instructions for accessing the data through the Hugging Face API, an overview of the dataset community on Hugging Face, and an analysis of the dataset cards.



	dataset_name	author	dataset_creation_time	downloads	has_card	has_nonempty_card	task	domain
0	super_glue	huggingface	Tue Jan 25 16:34:18 2022 +0100	1403269.0	True	True	text-classification,token-classification,quest...	nlp
1	glue	huggingface	Tue Jan 25 16:34:03 2022 +0100	1140355.0	True	True	text-classification	nlp
...	...	...	...	...	...	...	...	...
24063	fffhyyhh666/Mouth-64	fffhyyhh666	Mon Aug 29 14:52:42 2022 +0000	0.0	False	False	None	None
24064	IronDice/esdeath	IronDice	Fri Mar 10 21:46:51 2023 +0000	0.0	False	False	None	None

24065 rows × 8 columns

	dataset_name	author	dataset_creation_time	downloads	task	domain	dataset_card	total_word_cnt	follow_template
0	super_glue	huggingface	Tue Jan 25 16:34:18 2022 +0100	1403269.0	text-classification,token-classification,quest...	nlp	---\nannotations_creators:\n-expert-generated...	517.0	1.0
1	glue	huggingface	Tue Jan 25 16:34:03 2022 +0100	1140355.0	text-classification	nlp	---\nannotations_creators:\n-other\nlanguage_...	1388.0	1.0
...	...	...	...	...	...	...	...	...	...
7431	irds/mmarco_v2_vi_train	irds	Thu Jan 5 03:29:58 2023 +0000	0.0	text-retrieval	None	---\npretty_name: '`mmarco/v2/vi/train`'\nview...	74.0	0.0
7432	autoevaluate/autoeval-staging-eval-project-976...	autoevaluate	Thu Jul 21 15:35:27 2022 +0000	0.0	None	None	---\ntype: predictions\ntags:\n-autotrain\n- ...	48.0	0.0

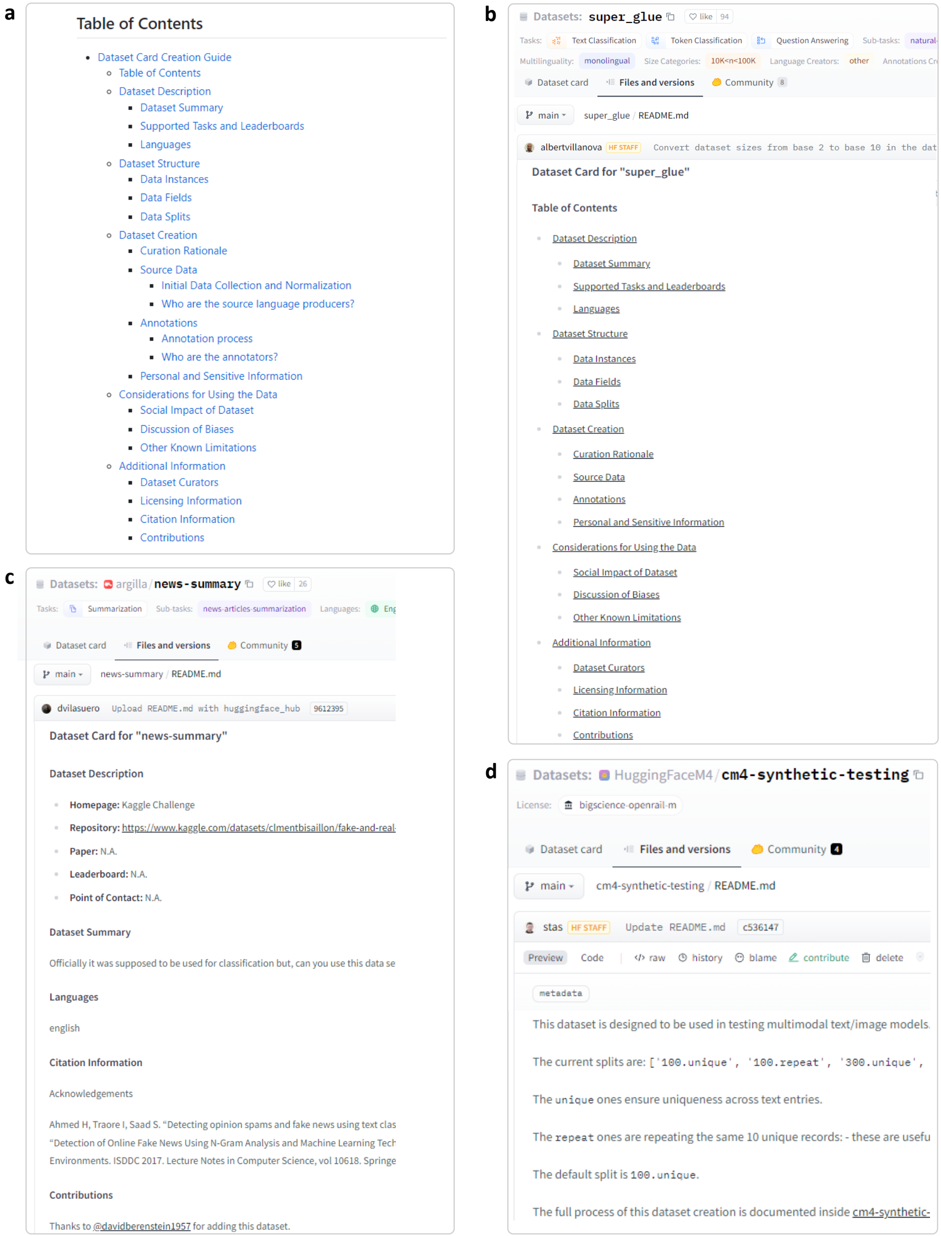
7433 rows × 9 columns

	dataset description						dataset structure	
	has_section	section_length_proportion	subsection_title	section_content	word_cnt	not_empty	has_section	section_leng
super_glue	1	0.268182	Dataset Summary	Dataset Description\nHomepage: https://github....	118	1	1	
glue	1	0.712919	Supported Tasks and Leaderboards;Languages;Dat...	Dataset Description\nHomepage: https://nyu-mll...	894	1	1	
...	...	...	...	...	...	...	...	
irds/mmarco_v2_vi_train	0	0.000000	None	None	0	0	0	

7433 rows × 36 columns

Figure S1: **Metadata Provided by the Repository for the Datasets and Dataset Cards.** (**a**) *Metadata for the Datasets:* The $dataset\_info.parquet$ in the $Data$ folder stores the metadata we extracted of the 24,065 datasets as of Mar 16th, 2023. The metadata include the creation time, author, downloads, whether the dataset has a (non-empty) dataset card, the task category, and the task domain of the dataset. (**b**) *Metadata for the Datasets Cards:* The $datasetcard\_info.parquet$ in the $Data$ folder stores the information we extracted of the 7,433 dataset cards. The information include the dataset name, author, creation time, number of downloads, task category, task domain, content ot the dataset card, total word count, and whether the dataset card follows the template. (**c**) *Information about the Sections of the Dataset Cards:* The $datasetcard\_sections\_info.parquet$ in the $Data$ folder stores the information of the sections of the dataset cards. The sections include Dataset Description, Dataset Structure, Dataset Creation, Considerations for Using the Data, Additional Information. For each section, we provide whether a dataset card has this section (and whether it's empty), the subsections of the section, section length proportion of the section, the content of the section, and the word count of the section.

B Illustrations for Dataset Cards Suggested by Hugging Face Community



a

Table of Contents

- Dataset Card Creation Guide
  - Table of Contents
  - Dataset Description
    - Dataset Summary
    - Supported Tasks and Leaderboards
    - Languages
  - Dataset Structure
    - Data Instances
    - Data Fields
    - Data Splits
  - Dataset Creation
    - Curation Rationale
    - Source Data
      - Initial Data Collection and Normalization
      - Who are the source language producers?
    - Annotations
      - Annotation process
      - Who are the annotators?
    - Personal and Sensitive Information
  - Considerations for Using the Data
    - Social Impact of Dataset
    - Discussion of Biases
    - Other Known Limitations
  - Additional Information
    - Dataset Curators
    - Licensing Information
    - Citation Information
    - Contributions

b

Convert dataset sizes from base 2 to base 10 in the dat

Dataset Card for "super_glue"

Table of Contents

- Dataset Description
  - Dataset Summary
  - Supported Tasks and Leaderboards
  - Languages
- Dataset Structure
  - Data Instances
  - Data Fields
  - Data Splits
- Dataset Creation
  - Curation Rationale
  - Source Data
  - Annotations
  - Personal and Sensitive Information
- Considerations for Using the Data
  - Social Impact of Dataset
  - Discussion of Biases
  - Other Known Limitations
- Additional Information
  - Dataset Curators
  - Licensing Information
  - Citation Information
  - Contributions

c

Upload README.md with huggingface_hub

Dataset Card for "news-summary"

Dataset Description

- **Homepage:** Kaggle Challenge
- **Repository:** https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-
- **Paper:** N.A.
- **Leaderboard:** N.A.
- **Point of Contact:** N.A.

Dataset Summary

Officially it was supposed to be used for classification but, can you use this data se

Languages

english

Citation Information

Acknowledgements

Ahmed H, Traore I, Saad S. "Detecting opinion spams and fake news using text clas
"Detection of Online Fake News Using N-Gram Analysis and Machine Learning Tech
Environments. ISDDC 2017. Lecture Notes in Computer Science, vol 10618. Springe

Contributions

Thanks to @davidberenstein1957 for adding this dataset.

d

Update README.md

This dataset is designed to be used in testing multimodal text/image models.

The current splits are: ['100.unique', '100.repeat', '300.unique',

The unique ones ensure uniqueness across text entries.

The repeat ones are repeating the same 10 unique records: - these are usefu

The default split is 100.unique.

The full process of this dataset creation is documented inside cm4-synthetic-

Figure S2: **Illustration of Adherence to Community-Endorsed Dataset Card.** (**a**) *Community-Endorsed Dataset Card Struture:* Hugging Face community provides a suggested dataset card structure, which contains five main sections: *Dataset Description*, *Dataset Structure*, *Dataset Creation*, *Considerations for Using the Data*, and *Additional Information*. (**b**) *Example of a Dataset Card Conforming to the Community Guidelines:* A dataset card is considered to conform to the community guidelines when it includes the five main sections outlined in the community guidelines, with the corresponding content provided for each section. (**c**) *Example of Dataset Cards Not Following Community Guidelines (1):* A dataset card is considered non-conforming if it omits any of the five main sections provided in the suggested dataset card structure. (**d**) *Example of Dataset Cards Not Following Community Guidelines (2):* This dataset card contains only a few words and does not follow the structure at all.

C Method

C.1 Accessing and Parsing Dataset Cards

In this work, we analyze datasets hosted on Hugging Face, a popular platform that provides a wealth of tools and resources for AI developers. One of its key features is the Hugging Face Hub API, which grants access to a large library of pre-trained models and datasets for various tasks. With this API, we obtained all 24,065 datasets hosted on the Hub as of March 16th, 2023.

Dataset cards are Markdown files that serve as the README for a dataset repository. They provide information about the dataset and are displayed on the dataset's homepage. We downloaded all dataset repositories hosted on Hugging Face and extracted its README file to get the dataset cards. For further analysis of the documentation content, we utilized the Python package mistune (`https://mistune.readthedocs.io/en/latest/`) to parse the README file and extract the intended content. The structure of dataset cards typically consists of five sections: *Dataset Description*, *Dataset Structure*, *Dataset Creation*, *Additional Information*, and *Considerations for Using the Data*, as recommended by Hugging Face community. Examples of dataset cards, as shown in **Fig. S2**, illustrate the essential components and information provided by dataset cards. We identified and extracted different types of sections through parsing and word matching of the section heading.

C.2 Human-Annotated Dataset Card Evaluation Methodology and Criteria

We conducted an evaluation on a sample of 150 dataset cards from a total of 7,433. The assessment involved five human annotators and focused on seven key aspects of the dataset cards:

- **Structural Organization:** How well is the documentation structured with headings, sections, or subsections?
- **Content Comprehensiveness:** How comprehensive is the information provided in the documentation?
- **Dataset Description:** How effectively does the documentation describe the dataset?
- **Dataset Structure:** How well does the documentation explain the underlying data structure of the dataset?
- **Dataset Preprocessing:** How well does the documentation describe any preprocessing steps applied to the data?
- **Usage Guidance:** How well does the documentation offer guidance on using the dataset?
- **Additional Information:** How well does the documentation provide extra details such as citations and references?

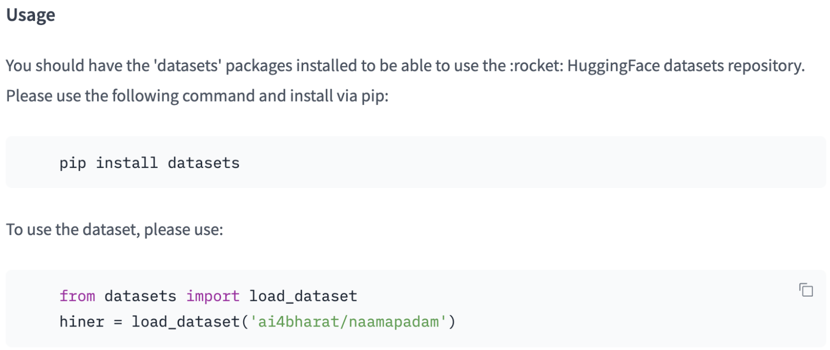
Each aspect received a score on a scale from 0 to 5, with the following score metrics:

Score	Description
5	Exceptionally comprehensive and effective
4	Very good and thorough
3	Moderately satisfactory
2	Insufficient
1	Poor and inadequate
0	Absent

Table S1: **Metrics of the Scores**

D Additional Analysis of *Usage* Section

Among 7,433 dataset cards, there are 567 dataset cards uploaded by 52 distinct practitioners that contain a *Usage* section, instructing how to use the dataset through text and codes. A specific example of *Usage* section is from ai4bharat/naamapadam, which has 469 downloads and has a *Usage* section to instruct how to use the dataset (**Fig. S3**).



Usage

You should have the 'datasets' packages installed to be able to use the :rocket: HuggingFace datasets repository. Please use the following command and install via pip:

```
pip install datasets
```

To use the dataset, please use:

```
from datasets import load_dataset
hiner = load_dataset('ai4bharat/naamapadam')
```

Figure S3: **Example of a *Usage* Section**

Intuitively, a *Usage* section could give users quick instructions on how to use the dataset, which could make the dataset more accessible, transparent, and reproducible. To verify this intuition, we conduct an experiment to quantify how the *Usage* section will affect the dataset's popularity.

In our experiment, we trained a BERT [11] Model using the content of dataset cards and their corresponding download counts. To ensure comparability, the download counts were normalized to a range of [0,1] and stratified monthly based on the dataset's creation time. This ranking system assigned a rank of 1 to the dataset with the highest downloads within a given month, and a rank of 0 to the dataset with the lowest downloads.

Using the dataset card content, the trained BERT Model predicted the download counts. Subsequently, we conducted a test using 567 dataset cards that included a *Usage* section. For this test, we deliberately removed the *Usage* section from the dataset cards and employed the BERT Model to predict the download counts for these modified cards. The resulting predictions are summarized in the table below:

	Predicted Score of Downloads
Dataset Card with Usage Section	0.3917
Remove the Usage Section	0.3732
Reduction upon Removal	**-0.0185**

Table S2: **Impact of Usage Section on Predicted Score of Downloads**

The average predicted score of downloads after removing the *Usage* section is 0.0185 lower compared to the original dataset card. This indicates a decrease in the number of downloads, highlighting the negative impact of not including a *Usage* section.

In future research, it would be valuable to further investigate the effect of adding a *Usage* section to the dataset cards that do not have one originally. A randomized controlled trial (RCT) experiment could be conducted to assess whether the inclusion of a *Usage* section leads to an increase in downloads.

E Optional Metrics for Datasets

In our analysis, we employ downloads as a metric to gauge the popularity of the dataset. Numerous factors can influence the download count, including the dataset's publication date and its associated research field. Moreover, aside from dataset downloads, we can incorporate other indicators of dataset popularity, such as the count of models utilizing the datasets and the corresponding download counts.

To address the concerns of factors that might affect downloads, we expanded our dataset analysis by extracting more metadata from the Hugging Face dataset information. We collected data such as the models utilizing the corresponding dataset, the total number of downloads for these models, and the dataset's task domain. The primary dataset tasks recognized by Hugging Face encompass Multimodal, Computer Vision, Natural Language Processing, Audio, Tabular and Reinforcement Learning. Among the total of 7,433 dataset cards, 1,988 are categorized as NLP dataset cards, 198 are related to computer vision, and 102 pertain to multimodal datasets. We proceeded with additional analysis by employing the following metrics:

1. We integrated dataset downloads with the downloads of models employing the dataset, which can be termed as *"secondary usage of the dataset"*.
2. Task domains were specified.
3. A time range (measured in months) was selected, encompassing dataset cards created within the designated time frame and domain.
4. Selected dataset cards were ranked within each domain for each time range and then normalized to a range of $[0, 1]$.

By adopting this approach, we account for the dataset's publication time, task domain, secondary dataset usage, as well as the number of downloads. We conducted a word count analysis using this new metric and attained results consistent with our prior analysis that datasets with higher rankings tend to have longer dataset cards, as shown in **Fig. S4**.

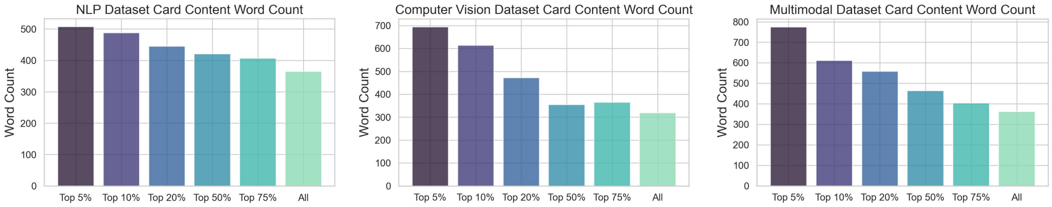


Figure S4: **Length Correlates with Dataset Quality.** In the updated metrics, there's a notable trend where higher-ranked dataset cards tend to be longer. This suggests that these dataset cards encompass more comprehensive and detailed information.

The finding enables us to contemplate an alternative metric option, factoring in publication time, research area, and secondary dataset usage. However, the results remain aligned with our previous analysis, which solely considered download counts, highlighting the reasonableness of using download counts as metrics.

F Additional Analysis of Each Section in the Dataset Card

Section. 5 offers a concise summary of each section, complemented by topic modeling results for the most engaging section, *Considerations for Using the Data*. In addition, **Table. 1** provides a clear presentation of the community-endorsed dataset card, including suggested sections, subsections, and their corresponding descriptions. The completion rates of subsections within each section are depicted in **Fig. 4**, which suggests a general adherence to the community-endorsed dataset card. In the subsequent paragraph, a comprehensive analysis of each section is provided, offering further insight into the content covered.

Dataset Description The *Dataset Description* section contains the fundamental information about a dataset, and is comprised of three subsections: *Dataset Summary*, *Supported Tasks and Leaderboards*, and *Languages*. As depicted in **Fig. 4**, *Dataset Summary* is the most frequently filled-out subsection in the *Dataset Description* section, with a filled-out rate of 94.5% and 80.0% in the top 100 downloaded dataset cards and all 7,433 dataset cards, respectively. This underscores the importance of providing a brief summary of the dataset, which can enhance its accessibility to users and, in turn, promote its use. On the other hand, the finer-grained subsections of *Dataset Description*, such as *Supported Tasks and Leaderboards* and *Languages*, have a relatively low filled-out rate. This may be due to the fact that people tend to provide only a brief mention of this information in the *Dataset Summary* section, instead of elaborating on it in a separate section. However, separating this information into distinct subsections can help to emphasize its importance. Given that tasks and languages are essential features of a dataset, it could be better for developers to follow the guidelines and write the information in the corresponding sections.

Dataset Structure Overall, dataset cards conform well to the official guidelines in the *Dataset Structure* section, particularly in the case of the top 100 downloaded dataset cards. Specifically, 95.3% of the top 100 downloaded dataset cards contain *Data Instances* in the *Dataset Structure* section, 98.8% of them contain *Data Fields*, and 97.7% of them contain *Data Splits*. The *Dataset Structure* section offers detailed information about the dataset's composition, with *Data Instances* providing examples and descriptions of typical instances in the dataset, *Data Fields* describing the fields present in the dataset, and *Data Splits* providing information about the criteria for splitting the data, as well as the size and name of each split. The high filled-out rate of these subsections highlights their importance and serves as an example for practitioners to follow when providing information about the *Dataset Structure*.

Dataset Creation *Dataset Creation* encompasses both technical and ethical considerations. Technical aspects, such as *Source Data*, which provides information about the initial data collection and normalization, and the source language producers, have the highest filled-out rate, at 70.8% and 70.6% for all datasets and the top 100 downloaded datasets, respectively. The *Annotations* subsection, which includes information about the annotation process and annotators, receives moderate attention, with a filled-out rate of 59.5% and 52.8% for all dataset cards and the top 100 downloaded dataset cards, respectively. Subjective issues, such as *Curation Rationale*, which outlines the motivation and reasons behind dataset curation, are included in 55.8% of dataset cards within the *Dataset Creation* section. Notably, the *Personal and Sensitive Information* subsection has a low filled-out rate, with only 35.3% of dataset cards discussing it in the *Dataset Creation* section. This is understandable, as limited datasets contain sensitive data that reveals information such as racial or ethnic origins, religious beliefs, political opinions, and so on. Nevertheless, this subsection is indispensable, as it helps ensure that the dataset is being handled ethically and in compliance with relevant regulations and laws. By providing information about any personal or sensitive data in the dataset, researchers and data scientists can take appropriate measures to protect the privacy and security of individuals represented in the data.

Considerations for Using the Data **Section. 4** highlights that *Considerations for Using the Data* is the section of a dataset card that receives the lowest attention. However, despite this, three prominent topics discussed in this section have been identified by the community: *Social Impact of Dataset*, *Discussion of Biases*, and *Other Known Limitations*. These topics are prevalent among both the entire set of 7,433 dataset cards and the top 100 downloaded dataset cards, all have a filled-out rate larger than 50%. Specifically, 80.0% of the top 100 downloaded dataset cards that include *Considerations for Using the Data* discuss the *Social Impact of Dataset*, describing the potential ways that the dataset may impact society. For example, the datasets for evaluating the fairness of pre-trained legal language models and techniques [8] states the following sentence in its *Social Impact of Dataset* section: "This work can help practitioners to build assisting technology for legal professionals with respect to the legal framework (jurisdiction) they operate." Additionally, 73.3% of the top 100 downloaded dataset cards discuss the biases of the dataset, such as biases of the data distribution or data collection process. (e.g. "This dataset is imbalanced"; "Since the data is from human annotators, there are likely to be biases.") The *Other Known Limitations* subsection outlines other limitations of the dataset, such as annotation artifacts, and is present in 57.2% of the *Considerations for Using the Data* sections. This subsection is important because it helps potential users understand the potential limitations and drawbacks of the dataset, which can inform their decision-making process when selecting a dataset for their research.

Overall, the high filled-out rate of the subsections of *Considerations for Using the Data* underscores the importance of considering the potential biases and limitations of a dataset, as well as its potential impact on society, when selecting and using a dataset for research purposes, and suggests researchers and data scientists are increasingly put more emphasis on the ethical and technical implications of their work.

Additional Information The *Additional Information* section of the dataset card includes details about the dataset curators, licensing information, citation information, and contributions. Our analysis shows a high rate of completion for citation information and contributions among the top 100 downloaded dataset cards that include this section. Of the top 100 downloaded dataset cards that contain *Additional Information*, 95.6% include the *Contributions* section, which typically acknowledges contributors with a statement like "Thanks to @github-username for adding this dataset", as suggested by the community-endorsed dataset card. Additionally, 94.5% of these dataset cards include citation information in BibTex format.

These findings emphasize the importance that researchers place on community sharing and recognition of contributions. Such emphasis can promote a healthy community ecosystem for sharing and discussing ideas and therefore prompt the development of the research field.

Other The *Other* section in a dataset card includes topics that are not covered by the five sections of the community-endorsed dataset card. Our analysis identifies two prominent topics that people discuss in this section. The first is *About*, which is similar to the *Dataset Description* section and accounts for 16.6% of *Other* sections. The second is *Usage*, which has a 33.2% filled-out rate of all discussions in the *Other* section. Indeed, the *Usage* section in a dataset card is important because it could provide users with information on how to use the dataset, including instructions on how to download and access the data, as well as how to preprocess or transform the data for various use cases. A clear and detailed *Usage* section can help users avoid common pitfalls or errors, saving time and effort for researchers and developers who are using the dataset for their projects. This, in turn, increases the reproducibility, transparency, and usage of the dataset. We suggest that dataset creators include a comprehensive *Usage* section in their dataset card to facilitate the use and reproducibility of the dataset. Furthermore, we recommend that the community incorporates this key information into their suggested dataset card to better serve the needs of the community.